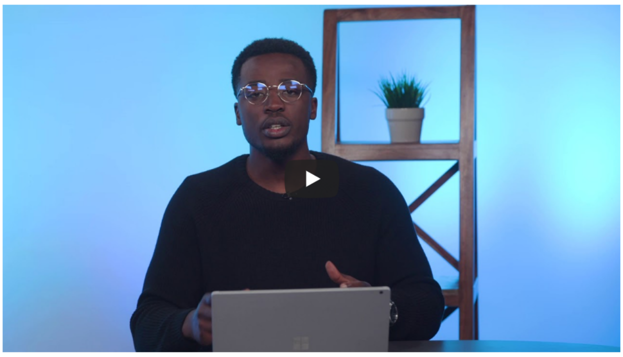
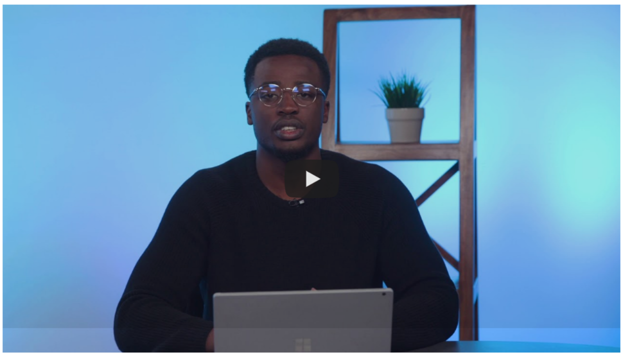
Operationalizing Models

After you have trained your machine learning model and evaluated it to the point where you are ready to use it outside your own development or test environment, you need to deploy it somewhere. Another term for this is **operationalization**.



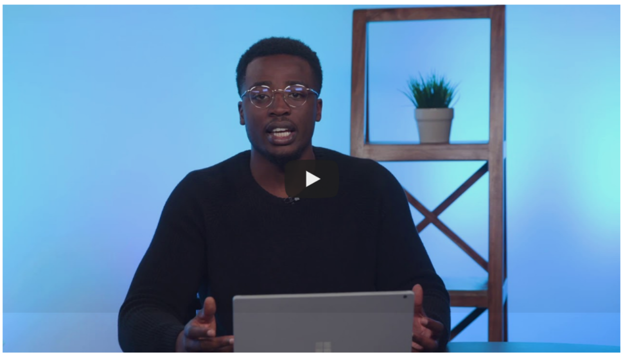
Real-time Inferencing

The model training process can be very compute-intensive, with training times that can potentially spann across many hours, days, or even weeks. A trained model, on the other hand, is used to make decisions on new data quickly. In other words, it infers things about new data it is given based on its training. Making these decisions on new data on-demand is called **real-time inferencing.**



Batch Inferencing

Unlike real-time inferencing, which makes predictions on data as it is received, **batch inferencing** is run on large quantities (batches) of existing data. Typically, batch inferencing is run on a recurring schedule against data stored in a database or other data store.



QUIZ QUESTION

When you deploy a Machine Learning model used for real-time scoring, or batch inferencing, there are several steps you must follow. Select the *required* steps in the list below.

(Select all that apply.)

- ✓ Save and retrieve the model file in any format
- □ Create a metadata file that describes the model
- □ Create a training script
- ✓ Create a scoring script
- □ Create a schema file that describes the web service input
- ✓ Create a real-time scoring web service

SUBMIT

NEXT